Submitted to Medinfo 2010.
draft material -- please do not cite

Exploiting UMLS Semantics for Quality Assurance Purposes

Halit Erdogan[a], Esra Erdem[a], Olivier Bodenreider[b]

[a] *Faculty of Engineering and Natural Sciences, Sabancı University, Turkey*
[b] *US National Library of Medicine, NIH, Bethesda, USA*

Abstract

Objectives*: To quantify semantic inconsistency in UMLS concepts from the perspective of their hierarchical relations and to show through examples how semantically-inconsistent concepts can help reveal erroneous synonymy relations.* ***Methods****: Inconsistency is defined in reference to concepts from the UMLS Metathesaurus. Consistency is evaluated by comparing the semantic groups of the two concepts in each pair of hierarchically-related concepts. A limited number of inconsistent concepts was inspected manually.* ***Results****: 81,512 concepts are inconsistent due to the differences in semantic groups between a concept and its parent. Four examples of wrong synonymy are presented.* ***Conclusions****: A vast majority of inconsistent hierarchical relations are not indicative of any errors. We discovered an interesting semantic pattern along hierarchies, which seems associated with wrong synonymy.*

Keywords:

Unified Medical Language System, Quality Assurance. Semantics.

Introduction

Capsule of adrenal gland is an anatomical concept found in the Foundational Model of Anatomy (FMA) and the NCI Thesaurus. In the FMA, it is defined as a subclass of *Capsule*. Once integrated in the Unified Medical Language System (UMLS), *Capsule of adrenal gland* (C1181304) appears as a child of the concept *capsule (pharmacologic)* (C1181304), for which "*Capsule*" is also a name. Of course, *Capsule* is an ambiguous name used by both anatomy and pharmacology specialists. In fact, a search for "capsule" in the UMLS yields 4 concepts (Table 1). Surprisingly, none of these concepts pertains to macroscopic anatomical structures.

The issue here is both the absence in the UMLS of a concept for the membranous layer surrounding an organ and the wrong association in the UMLS Metathesaurus of this meaning with the pharmacologic concept *capsule (pharmacologic)*.

We stumbled upon this error while exploring the UMLS Metathesaurus for creating lists of drug form terms in preparation for the i2b2 Challenge in Natural Language Processing for Clinical Data on Medication Extraction[1]. We were surprised to find anatomical concepts such as *Capsule of adrenal gland* in the descendants of the concept *Solid Dose Form* (C1378566).

Table 1. Four concepts for Capsule in the UMLS

Name	CUI	Sem. Type	Sem. Group
capsule (pharmacologic)	C0006935	Biomedical or Dental Material	CHEM
Capsule Dosing Unit	C1706433	Quantitative Concept	CONC
Capsule Shape	C1704652	Qualitative Concept	CONC
Microbial anatomical capsule structure	C1325531	Cell Component	ANAT

Quality assurance in biomedical terminologies is an active field of research [1]. Various research groups have investigated quality in the UMLS, addressing issues including terminological cycles [2], ambiguity of concepts [3, 4], concept categorization [3, 5]. Consistency across hierarchies has been addressed by [6], while [7] have studied the concistency of Metathesaurus relations against Semantic Network relations. More recently, the semantic groups have been used for analyzing the consistency of Metathesaurus relations [8]. Most closely related to our work is a study of the validity of concepts associated with multiple semantic groups [9]. Our current work uses the same approach to assess the validity, not of single concepts, but of pairs of hierarchically-related concepts.

The objective of this paper is to quantify semantic inconsistency in UMLS concepts from the perspective of their hierarchical relations and to show through examples how semantically inconsistent concepts can help reveal erroneous synonymy relations. The specific contribution of this paper is to leverage the semantic groups for identifying inconsistencies and to consider not only the semantics directly ascribed to a concept, but also the semantics it inherits from its ancestors.

[1] https://www.i2b2.org/NLP/Medication/

Background

The Unified Medical Language System® (UMLS®) [10] includes two sources of semantic information: the Metathesaurus® and the Semantic Network. The UMLS Metathesaurus was assembled by integrating some 150 source vocabularies. It contains more than 2 million concepts, i.e., clusters of synonymous terms coming from multiple source vocabularies identified by a Concept Unique Identifier (CUI). More than 36 million relations are recorded between these concepts. Several types of relationships among concepts are recorded in the Metathesaurus: parent / child of (PAR / CHD) and broader / narrower than (RB / RN) essentially correspond to hierarchical relations, while the other relationships are associative. More than 7.5 million hierarchical relations are represented in the Metathesaurus.

The Semantic Network is a much smaller network of 135 Semantic Types (STs) organized in a tree structure [11]. Each Metathesaurus concept is assigned at least one ST. Groupings of STs, called semantic groups (SGs), represent subdomains of biomedicine such as *Anatomy*, *Chemicals & Drugs*, and *Disorders* [12]. Each ST belongs to one and only one SG.

Methods

Computation

We say that a UMLS concept is *inconsistent* if the following two conditions hold for the concept:

- it belongs to two different semantic groups (except CONC) either directly or via its ancestors;
- it does not have any inconsistent ancestor (i.e., the inconsistency of the concept is not due to inheritance, it is original).

Therefore, to compute all inconsistent concepts, we need some method to find all the ancestors of a concept and their semantic groups, and check the inconsistency of each ancestor.

A naive method to find all the ancestors of a concept and their semantic groups can be described as follows:

1. for each concept, compute all its ancestors;
2. for each ancestor, check that it is not inconsistent;
3. if no ancestor is inconsistent then check for each pair of ancestors whether they belong to different groups; else quit.

However, there are over 2,000,000 concepts in the Metathesaurus and some of them have too many ancestors (over 800); and thus it may not be practical to generate all of them at Step 1 above. Note that, to check the inconsistency of each ancestor at Step 2, we can apply the three-stage method above recursively: for each ancestor, find its ancestors and their semantic groups, and check the inconsistency of its ancestors, and so on.

Considering that a slight modification of the UMLS graph or a slight modification of the definition of inconsistency requires a new sequence of computations from the very beginning (that is, we may not reuse the previous results), we need to find more efficient methods that do not require programming efforts from the user and that can handle recursion.

We introduce a new method for computing inconsistent concepts: first, we divide the set of all UMLS concepts into smaller sets (e.g., of size 20,000); then, for each set, we compute in parallel all ancestors of its elements in the whole UMLS graph that are not inconsistent. Therefore, we compute the inconsistent concepts, as we compute their ancestors and check their inconsistency.

We have realized this method using a computational methodology, called Answer Set Programming (ASP) [13, 14]. The idea is to define the ancestors of concepts, and the inconsistent concepts by means of (possibly recursive) rules, and then call an existing ASP system (e.g., the ASP solver CLASP) to find inconsistencies based on these definitions. Figure 1 shows a sample ASP definition of inconsistency for a set of concepts.

Evaluation

One of us (OB) performed a detailed review of 200 inconsistencies involving the semantic groups *Anatomy* and *Chemicals & Drugs*, groups in which the "capsule" error was originally observed. In addition, we performed a casual inspection of the four major groups of inconsistencies in order to identify categories of inconsistencies.

Results

Quantitative results

We identified 334,396 inconsistent concepts. Out of these concepts, 81,512 concepts are inconsistent due to the following reason: the semantic group of the parent differs from that of the source concept, and no ancestor of the concept is inconsistent. For example, the concept *Anti-purkinje cell antibody* (C0443893) is one of these 81,512 concepts: its semantic group is *Chemicals & Drugs*, whereas its parent *Purkinje Cells* (C0034143) belongs to the semantic group *Anatomy*; furthermore, no ancestor of *Anti-purkinje cell antibody* is inconsistent.

The distribution of the number of inconsistencies by semantic group of the source concept is listed in Table 2. Two semantic groups, *Disorders* and *Physiology*, represent less than 25% of all UMLS concepts, but concentrate 80% of the inconsistencies. This map of inconsistencies can be further refined by looking at the semantic group of the parent of inconsistent concepts in reference to that of the source concept. The number of inconsistent *child_of*[2] relations by semantic group of the source and parent concepts is listed in Table 3. For example, 10,732 inconsistent *child_of* relations involve a concept from

[2] We use a generic *child_of* relationship to represent the various kinds of hierarchical relations in the Metathesaurus (child and narrower than).

the semantic group *Disorders* as the child and a concept from the semantic group *Anatomy* as the parent. Note that the number of inconsistent *child_of* relations is slightly higher than the number of inconsistent concepts, since a given concept can be involved in several inconsistent *child_of* relations.

Table 2. Distribution of the number of inconsistencies by semantic group of the source concept

	SG (source)	# conc	%
ACTI	Activities & Behaviors	750	1%
ANAT	Anatomy	809	1%
CHEM	Chemicals & Drugs	3482	4%
CONC	Concepts & Ideas	--	
DEVI	Devices	2463	3%
DISO	Disorders	30704	38%
GENE	Genes & Molecular Sequences	158	0%
GEOG	Geographic Areas	22	0%
LIVB	Living Beings	995	1%
OBJC	Objects	1084	1%
OCCU	Occupations	134	0%
ORGA	Organizations	236	0%
PHEN	Phenomena	4257	5%
PHYS	Physiology	34366	42%
PROC	Procedures	2052	3%
	Total	*81512*	*100%*

Wrong synonymy relations

Including the "capsule" error, four errors of the same type were identified by manual review of pairs of concepts with inconsistent *child_of* relations, two of which involve the semantic groups *Anatomy* and *Chemicals & Drugs*.

Capsule. Parents of *capsule (pharmacologic)* (C1181304) from the semantic group *Chemicals & Drugs*, presented in the introduction, include anatomical concepts such as *Membranous layer* (C2338391), as well as drug concepts (e.g., *Pill* (C0994475)). Analogously, mixed semantic is found among its children, with anatomical concepts such as *Capsule of adrenal gland* (C1181304) and drug concepts including *Oral Capsule* (C0991533). In order to address the wrong synonymy in *capsule (pharmacologic)*, a distinct concept should be created for the anatomical capsule, with a semantic type from the semantic group *Anatomy*.

Retina / Retinol. Two types of parents can be observed for the concept *Retina* (C0035298). On the one hand, there are concepts from the semantic group *Chemicals & Drugs*, such as *All-Trans-Retinol* (C0087161) and *Aldehydes* (C0001992). On the other, there are concepts from the semantic group *Anatomy*, including *Wall of eyeball* (C0929391). Except for some lexical resemblance, it is unclear what caused this error. However, there seems to be a wrong synonymy issue, because three of the children of *Retina* are also from the semantic group *Chemicals & Drugs* (e.g., *Retinal | bld-ser-plas* (C1972646)), while most children are anatomical concepts (e.g., *Retinal Neurons* (C2350331)).

Two additional examples of wrong synonymy involving other semantic groups than *Anatomy* and *Chemicals & Drugs* were identified in this study.

California plant / state. Parents of the concept *California* (C0006754) from the semantic group *Geographic Areas* include *Pacific States* (C0524818) from the same group and *Geraniaceae* (C0996910) from the semantic group *Living Beings*. (*Geraniaceae* is a family of plants comprising, among others, Geranium). Analogously, the children of *California* include other geographic areas (e.g., *San Francisco* (C0036152)), as well as plants (e.g., *California macrophylla* (C1891810)). This is another example of wrong synonymy. A distinct concept should be created for the plant family *Geraniaceae*, with a semantic type of Plant, from the semantic group *Living Beings*.

Transdermal / Skin Patch. The concept *Transdermal Patch* (C0991556) from the semantic group *Chemicals & Drugs* has two types of parents. On the one hand, there are several drug concepts, including *Patch drug form* (C0994894). On the other, parents such as *Lesion* (C0221198) belong to the semantic group *Disorders*. Moreover, the same mixed semantics can be observed among the children of *Transdermal Patch*, including the drug *Nicotine patches* (C0358855) and the clinical finding *Café-au-Lait Spots* (C0221263). The issue here is probably wrong synonymy related to "patch". A distinct concept should be created for the clinical finding *Skin Patch*, with a semantic type from the semantic group *Disorders*.

Other types of inconsistencies

In addition to wrong synonymy, several other types of inconsistencies were observed. In contrast to wrong synonymy, the other inconsistencies could be expected, since the UMLS is **not** an ontology of biomedicine, but rather a terminology integration system, which, by design, does not impose a semantic model to the terminologies it integrates [15].

Many terminologies such as the Medical Subject Headings (MeSH) organize their concepts for a particular purpose rather than based on ontological principles. For example, the hierarchical organization of MeSH descriptors supports information retrieval. In other words, MeSH hierarchical relations often reflect "aboutness" rather than subsumption [16]. One such example is the relation from the Japanese version of MeSH between *Orphan Drugs* (*Chemicals & Drugs*) and *Drug Industry* (*Organizations*).

The absence of formal distinction between types and roles in the semantic network can lead to inconsistent usage. For example, *Sheep milk* is considered food (*Objects*), while its parent concept *Milk* is considered a body substance (*Anatomy*).

We also found cases where concept categorization (i.e., the semantic type assigned to the concept) is arguable or inconsistent across similar concepts. One such example is *Animal Disease Models*, categorized as a disease model (*Disorders*), while its parent concept *Animal Model* is (wrongly) categorized as an animal (*Living Beings*.).

Finally, differences in granularity between hierarchically-related concepts are at the origin of some of the inconsistencies. For example, *Iodothyroglobulin* (*Chemicals & Drugs*) and *Thyroid colloid* (*Anatomy*) are definitely related, but their relationship – between molecular and macroscopic structures – is mereological in nature (part-whole) rather than taxonomic (is a).

Our casual review of four large sets of inconsistencies revealed that inconsistencies in these groups were essentially homogeneous within a group. We examined inconsistencies from the pairs of semantic groups with the largest number of inconsistent relations (10-20,000 per group). These are DISO-ANAT, DISO-PHYS, PHYS-PROC and PHYS-CHEM.

Most inconsistencies from **DISO-ANAT** come from the integration of the clinical synopsis from OMIM in the UMLS. In a clinical synopsis, disorders are grouped under anatomical structures (e.g., *Mouth Neoplasms* under *Oral cavity*). The corresponding relations, i.e., *Mouth Neoplasms* to *Oral cavity*, have been integrated as *child_of* relations in the UMLS, leading to this type of inconsistency.

Inconsistencies from DISO-PHYS generally correspond to pairs of hierarchically-related concepts in which the parent represents a quality being observed in a clinical observation (e.g., *Texture of hair*), categorized with a semantic type from the semantic group *Disorders* and the child one possible value for this quality (e.g., *Coarse hair*), categorized with a semantic type from the semantic group *Physiology*.

Inconsistencies from PHYS-PROC and PHYS-CHEM are related to the hierarchical organization of concepts in LOINC, i.e., how LOINC groups laboratory tests and clinical observations into classes, resulting in links between observations (from the semantic group *Physiology*) and chemical analytes (e.g., class of sodium plasma tests) or procedures (e.g., class of observations related to organ transplantation).

In these four groups, the large numbers of inconsistencies are generally not indicative of errors, but reflect the fact that, although used to form hierarchies, these relations are not hierarchical in nature (i.e., not is-a). Consistency checking based on the semantic groups assumes that the *child_of* relations are subsumption relations, which is not the case here.

Discussion

UMLS semantic framework

With its multiple layers, Metathesaurus concepts, semantic types from the Semantic Network and semantic groups, the UMLS provides a unique framework for checking semantic consistency. Semantic consistency between the Metathesaurus and the Semantic Network requires valid relations between concepts, valid relations between semantic types and accurate categorization of the concepts with semantic types. In contrast, inconsistency is indicative of a problem with at least one of these elements, but further analysis is required to pinpoint the problem causing the inconsistency. Additionally, semantic consistency between the Metathesaurus and the semantic groups (through the semantic types) is predicated upon the validity of the disjunction axioms added to the Semantic Network by the semantic groups.

As we have argued in the past [7], taking advantage of the UMLS semantic framework in the editing environment used by the Metathesaurus editors would help expose semantic inconsistencies at the time of editing and would likely reduce the number of such inconsistencies in the Metathesaurus release. Using the semantic groups rather than the semantic types for consistency checking seems appropriate in the context of terminology integration (as opposed to ontology development).

Lessons learned

One of the lessons learned from this analysis is that the approach we propose for identifying inconsistencies lacks specificity. In fact, we showed that a vast majority of inconsistent hierarchical relations are not indicative of any errors, but simply reflect the use of hierarchical relations for knowledge organization purposes.

Interestingly, we discovered that the four instance of wrong synonymy we identified exhibit a pattern of "semantic rupture" along the hierarchical structure of the terminology. By semantic rupture, we mean that, along one hierarchy, the source concept belongs to a given semantic group, its parent concept does not, but one of the parents of the parent belongs to the same group as the source concept. For example, in the "capsule" example presented earlier, the source concept is *Capsule of adrenal gland* (*Anatomy*). Its parent is *capsule (pharmacologic)* (*Chemicals & Drugs*), one parent of which is *Membranous layer* (*Anatomy*). We hypothesize that such pattern of semantic rupture might be a good marker for wrong synonymy and we plan to test it in future work.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

[1] Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. J Biomed Inform 2009;42(3):407-11

[2] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57-61

[3] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51

[4] Cimino JJ. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus. Proc AMIA Symp 2001:120-4

[5] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med 2004;31(1):29-44

[6] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform 2003;36(6):450-61

[7] McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Dordrecht; Boston: Kluwer Academic Publishers; 2002. p. 181-198

[8] Vizenor LT, Bodenreider O, McCray AT. Auditing associative relations across two knowledge sources. J Biomed Inform 2009;42(3):426-39

[9] Mougin F, Bodenreider O, Burgun A. Analyzing polysemous concepts from a clinical perspective: application to auditing concept categorization in the UMLS. J Biomed Inform 2009;42(3):440-51

[10] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91

[11] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics 2003;4(1):80-4

[12] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36(6):414-32

[13] Baral C. Knowledge Representation, Reasoning and Declarative Problem Solving: Cambridge University Press; 2003

[14] Lifschitz V. What is answer set programming? In: Proceedings of AAAI 2008: MIT Press; 2008

[15] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1-2):193-201

[16] Nelson SJ, Johnston D, Humphreys BL. Relationships in Medical Subject Headings. In: Bean CA, Green R, editors. Relationships in the organization of knowledge. New York: Kluwer Academic Publishers; 2001. p. 171-184

Address for correspondence

Esra Erdem (esraerdem@sabanciuniv.edu)

Olivier Bodenreider (olivier@nlm.nih.gov)

Table 3. Number of inconsistent parent relations by semantic group of the source concepts (rows) and parent concepts (columns) (NB: The abbreviations of the semantic groups are defined in Table 2)

src\target	ACTI	ANAT	CHEM	DEVI	DISO	GENE	GEOG	LIVB	OBJC	OCCU	ORGA	PHEN	PHYS
ACTI	0	0	1	1	238	0	0	13	17	102	18	81	121
ANAT	0	0	175	2	264	20	0	16	157	14	0	15	120
CHEM	1	188	0	193	59	72	1	1618	484	15	1	39	177
DEVI	0	6	1443	0	6	2	0	1	693	10	56	3	0
DISO	485	10732	82	94	0	9	1	167	28	333	2	3008	2492
GENE	0	12	66	0	25	0	0	18	4	1	0	3	25
GEOG	2	0	0	0	0	0	0	3	14	0	0	1	0
LIVB	44	12	306	1	225	0	1	0	53	254	3	12	97
OBJC	21	36	480	434	19	0	17	24	0	21	6	24	0
OCCU	13	1	1	0	6	0	0	50	2	0	6	4	12
ORGA	16	0	0	0	1	0	4	172	3	18	0	0	0
PHEN	42	27	107	1	1118	1	2	122	14	73	1	0	322
PHYS	101	6804	20412	224	2590	760	0	650	83	7210	0	824	0

```
% define the ancestors C2 of a concept C1 in set n
descendant(C1,C2) :- chd(C1,C2), set(n,C1).
descendant(C1,C2) :-descendant(C1,C), childOf(C,C2).

% define the concepts C with some inconsistent ancestor C1
descendantOfInconsistent(C) :- descendant(C,C1), inconsistent(C1).

% identify the groups G (except conc) that a concept C belongs to,
% such that C is not a descendant of an inconsistent ancestor
groupOfConcept(C,G) :- hasCategory(C,T), hasGroup(T,G), set(n,C), G != conc.
groupOfConcept(C,G) :- not descendantOfInconsistent(C), descendant(C,C1), hasCategory(C1,T), hasGroup(T,G), G != conc.

% a concept is inconsistent if it belongs to two different groups G and G1
inconsistent(C) :- groupOfConcept(C,G), groupOfConcept(C,G1), G<G1.
```

Figure 1. Defining inconsistencies in ASP